The Zelterman Estimate of Population Size under Heterogeneity

Dankmar Böhning

Quantitative Biology and Applied Statistics, School of Biological Sciences
University of Reading

IBC, Dublin, 13–18 July 2008
Introduction

Some Applications

Solutions to the Population Size Problem

Some Recent Results on Zelterman Estimation
  How are Chao’s and Zelterman’s Estimator related?
  Zelterman as MLE
  Zelterman can be extended to case data
  Zelterman extended to higher counts
  Some simulation results
a population has $N$ units of which $n$ are identified by some mechanism (trap, register, police database, ...).
Formulation of the Problem

- probability of identifying an unit is \((1 - p_0)\)
- so that \(N = (1 - p_0)N + p_0 \cdot N = n + p_0 \cdot N\)
- and the **Horvitz-Thompson** estimator follows:

\[
\hat{N} = \frac{n}{1 - p_0}
\]

- usually an estimate of \(p_0\) is required
Introduction

Formulation of the Problem as Frequencies of Frequencies

A common setting for estimating \( p_0 \) is the **Frequencies of Frequencies** setting:

- the identifying mechanism provides a count \( Y \) of repeated identifications (w.r.t. to a reference period)
- leading to frequencies \( f_1, f_2, \ldots, f_m \) of the counts 1, 2, .., \( m \) (\( m \) is the largest observed count)
- zero counts are not observed: hence \( f_0 \) is unknown
- Recall that \( N = f_0 + n = f_0 + f_1 + f_2 + \ldots + f_m \), so that \( \hat{f}_0 \) leads to \( \hat{N} \)
Introduction

Some Applications

Solutions to the Population Size Problem

Some Recent Results on Zelterman Estimation
  How are Chao’s and Zelterman’s Estimator related?
  Zelterman as MLE
  Zelterman can be extended to case data
  Zelterman extended to higher counts
  Some simulation results
Application Areas

- Epidemiology and Medicine
- Biology and Agriculture
- Social Science and Criminology
- Research on Terrorism

- Intravenous drug users in L.A. county were entered into the California Drug Abuse Data System (CAL-DADS)
- The data below refer to the frequency distribution of the episode count per drug user in 1989

The frequency distribution of the episode count per drug user for the year 1989:

<table>
<thead>
<tr>
<th>$f_0$</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>$f_4$</th>
<th>$f_5$</th>
<th>$f_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-11,982</td>
<td>3,893</td>
<td>1,959</td>
<td>1,002</td>
<td>575</td>
<td>340</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$f_7$</th>
<th>$f_8$</th>
<th>$f_9$</th>
<th>$f_{10}$</th>
<th>$f_{11}$</th>
<th>$f_{12}$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>214</td>
<td>90</td>
<td>72</td>
<td>36</td>
<td>21</td>
<td>14</td>
<td>20,198</td>
</tr>
</tbody>
</table>
Del Rio Vilas’s Data on Estimating Hidden Scrapie in Great Britain 2005

- Sheep is kept in holdings in Great Britain (and elsewhere)
- The occurrence of scrapie is monitored in the Compulsory Scrapie Flocks Scheme (CSFS) summarizing abbatoir survey, stock survey and the statutory reporting of clinical cases
- CSFS established since 2004

The frequency distribution of the \textit{scrapie count within each holding} for the year 2005:

<table>
<thead>
<tr>
<th>(f_0)</th>
<th>(f_1)</th>
<th>(f_2)</th>
<th>(f_3)</th>
<th>(f_4)</th>
<th>(f_5)</th>
<th>(f_6)</th>
<th>(f_7)</th>
<th>(f_8)</th>
<th>(n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>84</td>
<td>15</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>118</td>
</tr>
</tbody>
</table>
Formulation of the Problem and the Idea for its Solution

Suppose we can find some model for the count probabilities

\[ p_j = p_j(\lambda) \]

then estimate \( \lambda \) by some method (truncated likelihood) and then use the model for \( p_0 \):

\[ \hat{N} = \frac{n}{1 - p_0(\hat{\lambda})} \]
Formulation of the Problem and the Idea for its Solution

Only to illustrate: Poisson model for the count probabilities

\[ p_j = p_j(\lambda) = \exp(-\lambda) \frac{\lambda^j}{j!} \]

then estimate \( \lambda \) and arrive at:

\[ \hat{N} = \frac{n}{1 - \hat{p}_0} = \frac{n}{1 - \exp(-\hat{\lambda})} \]
Formulation of the Problem and the Idea for its Solution

However: using a simple Poisson model for the count probabilities

\[ p_j = p_j(\lambda) = \exp(-\lambda) \lambda^j / j! \]

is not appropriate, since

- every unit is different
- there is population heterogeneity

so that more realistic

\[ p_j = p_j(\lambda) = \int_0^{\infty} \exp(-t) t^j / j! \lambda(t) dt \]

where \( \lambda(t) \) stands for the heterogeneity distribution of the Poisson parameter
instead of providing an estimate $\hat{\lambda}(t)$ by means of parametric or nonparametric mixture models interest is on two alternatives:


\[ p_j = \int_0^\infty \exp(-t) t^j / j! \lambda(t) dt \]

with unknown $\lambda(t)$ for $t > 0$. Then, by the Cauchy-Schwartz inequality:

\[ p_1^2 \leq p_0^2 p_2 \iff \frac{p_1^2}{2p_2} \leq p_0 \]

leads to Chao’s lower bound estimate

\[ \hat{f}_0 = \frac{f_1^2}{2f_2} \]

2. robust approach of Zelterman (1988, *JSPI*)
The Idea of Zelterman (1988)

- he noted that

\[ \lambda = \frac{\lambda^{j+1}}{\lambda^j} = (j + 1) \frac{\lambda^{j+1}/(j + 1)!}{\lambda^j/j!} \]

\[ \lambda = (j + 1) \frac{\text{Po}(j + 1; \lambda)}{\text{Po}(j; \lambda)} \]

- leading to the proposal

\[ \hat{\lambda}_j = (j + 1) \frac{f_{j+1}}{f_j} \]

- and in particular for \( j = 1 \)

\[ \hat{\lambda} = \hat{\lambda}_1 = 2 \frac{f_2}{f_1} \]
\[ \hat{\lambda} = 2 \frac{f_2}{f_1} \] is robust in the sense that

- it is not affected by any changes in counts larger than 2
- count distribution need only to behave like a Poisson for counts of 1 or 2
Introduction

Some Applications

Solutions to the Population Size Problem

Some Recent Results on Zelterman Estimation

How are Chao’s and Zelterman’s Estimator related?
Zelterman as MLE
Zelterman can be extended to case data
Zelterman extended to higher counts
Some simulation results
Zelterman larger than Chao? Carothers capture-recapture data on the number of taxis in Edinburgh (42 sampling occasions)
Zelterman larger than Chao?

\[ \hat{N}_Z = \frac{n}{1 - \exp(-\hat{\lambda})} = n + \frac{n}{\exp(\hat{\lambda}) - 1} \approx n + \frac{n}{1 + \hat{\lambda} + \frac{1}{2} \hat{\lambda}^2 - 1} \]

\[ = n + \frac{n}{\hat{\lambda} + \frac{1}{2} \hat{\lambda}^2} = n + \frac{n}{\frac{2f_2}{f_1} + \frac{1}{2} \left( \frac{2f_2}{f_1} \right)^2} = n + \left( \frac{f_1^2}{2f_2} \right) \frac{n}{f_1 + f_2} \]

\[ \geq n + \left( \frac{f_1^2}{2f_2} \right) = \hat{N}_C \]

\[ \text{yes, if } \hat{\lambda} \text{ is small} \] (Böhning and Brittain SJOS 2008)
Zelterman, Chao and simple Poisson MLE for the four data sets

<table>
<thead>
<tr>
<th>Example</th>
<th>$n$</th>
<th>MLE</th>
<th>Chao</th>
<th>Zelterman</th>
<th>$\frac{f_2}{f_1}$</th>
<th>$\frac{n}{f_1+f_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scrapie</td>
<td>118</td>
<td>188</td>
<td>353</td>
<td>393</td>
<td>0.18</td>
<td>1.19</td>
</tr>
<tr>
<td>Drug Use L.A.</td>
<td>20,198</td>
<td>26,425</td>
<td>38,637</td>
<td>42,268</td>
<td>0.33</td>
<td>1.27</td>
</tr>
</tbody>
</table>

MLE:

$$\hat{N} = \frac{n}{1 - p_0(\hat{\lambda})}$$

where $\hat{\lambda}$ is MLE under homogenous Poisson
Zelterman Estimation offers Flexibility

Zelterman estimate truncates all counts different from 1 or 2: write

\[ 1 - p = p_1 = \frac{\exp(-\lambda)\lambda}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2} = \frac{1}{1 + \lambda/2} \]

\[ p = p_2 = \frac{\exp(-\lambda)\lambda^2/2}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2} = \frac{\lambda/2}{1 + \lambda/2} \]

and consider associated **binomial** log-likelihood

\[ f_1 \log(p_1) + f_2 \log(p_2) = f_1 \log(1 - p) + f_2 \log(p) \]

which is maximized for \( \hat{p} = \hat{p}_2 = \frac{f_2}{f_1 + f_2} \), or

\[ \hat{\lambda} = \frac{2\hat{p}_2}{1 - \hat{p}_2} = \frac{2f_2}{f_1} \]
Zelterman Estimation offers Flexibility

A likelihood framework offers generalizations:

- (correct) variance estimate of the Zelterman estimator (Fisher information) (Böhning 2008, *Statistical Methodology*)
- extension of the estimator for **case data**
- incorporation of **covariates** (binomial logistic regression with log-link function to the Poisson parameter) (Böhning and van der Heijden 2008 *Ann. Appl. Statist.*)
- efficiency
Zelterman Estimation: Extension to Case Data

Table: Illustration of Case Data with Individual Recapture Counts

<table>
<thead>
<tr>
<th>Unit $i$</th>
<th>Count $y_i$</th>
<th>$\delta_i$</th>
<th>Sex$_i$</th>
<th>Age$_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Male</td>
<td>34</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>Male</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>Female</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>-</td>
<td>Male</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>Female</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>Female</td>
<td>26</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Zelterman Estimation offers Flexibility

Binomial likelihood for **grouped** data

\[ f_1 \log(1 - p) + f_2 \log(p) \]

becomes for **case** data

\[ \sum_i (1 - \delta_i) \log(1 - p) + \delta_i \log(p) \]

which becomes with covariate information on case \( i \)

\[ p_i = \frac{\exp(\beta^T x_i)}{1 + \exp(\beta^T x_i)} \]

a **logistic regression** model
Zelterman Estimation offers Flexibility

covariate information on case \( i \)

\[
p_i = \frac{\exp(\beta^T x_i)}{1 + \exp(\beta^T x_i)}
\]

compare with parameterization in capture probability \( \lambda \)

\[
p_i = \frac{\lambda_i/2}{1 + \lambda_i/2}
\]

it follows that

\[
\lambda_i = 2 \exp(\beta^T x_i)
\]

and the generalization of the Horvitz-Thompson estimator is

\[
\sum_{i=1}^{n} \frac{1}{1 - \exp(-2e^{\beta^T x_i})}
\]
Generalizing the Idea of Zelterman: Improving Efficiency

- not only

\[ \lambda = 2 \frac{P(2; \lambda)}{P(1; \lambda)} \]

- but also

\[ \lambda = \lambda \left( \frac{\lambda + \lambda^2/2!}{\lambda + \lambda^2/2!} \right) = \frac{2\lambda^2/2! + 3\lambda^3/3!}{\lambda + \lambda^2/2!} = \frac{2P(2; \lambda) + 3P(3; \lambda)}{P(1; \lambda) + P(2; \lambda)} \]
The Zelterman Estimate of Population Size under Heterogeneity

Some Recent Results on Zelterman Estimation

Zelterman extended to higher counts

\[ \lambda = 2 \frac{\text{Po}(2; \lambda)}{\text{Po}(1; \lambda)} = \frac{2\text{Po}(2; \lambda) + 3\text{Po}(3; \lambda)}{\text{Po}(1; \lambda) + \text{Po}(2; \lambda)} = \frac{2\text{Po}(2; \lambda) + 3\text{Po}(3; \lambda) + 4\text{Po}(4; \lambda)}{\text{Po}(1; \lambda) + \text{Po}(2; \lambda) + \text{Po}(3; \lambda)} = \ldots \]

leads to the proposal

\[ \hat{\lambda} = \hat{\lambda}_1 = 2 \frac{f_2}{f_1}, \hat{\lambda}_2 = \frac{2f_2 + 3f_3}{f_1 + f_2}, \hat{\lambda}_3 = \frac{2f_2 + 3f_3 + 4f_4}{f_1 + f_2 + f_3}, \ldots \]

\[ \hat{N}_i = \frac{n}{1 - \exp(-\hat{\lambda}_i)} \]
Simulation Experiment

- **Goal:** Compare $\hat{N}_j = \frac{n}{1 - \exp(-\hat{\lambda}_j)}$ and Chao’s estimator $\hat{N}_C = n + \frac{f_1^2}{2f_2}$

- **Count data:** $f_j$ arise from $0.5 \text{Po}(j; 0.5) + 0.5 \text{Po}(j; \lambda)$ for $
\lambda = 1, 2, \ldots, 7$ and $j = 0, 1, 2, \ldots$

- **Population size:** $N = f_0 + f_1 + \ldots = 100$

- $f_0$ is truncated

- $N$ estimated using $\hat{N}_1, \hat{N}_2, \hat{N}_3$ and $\hat{N}_C$
The Zelterman Estimate of Population Size under Heterogeneity

Some Recent Results on Zelterman Estimation

Some simulation results

![Graph showing RMSE for different estimators](image-url)
The Zelterman Estimate of Population Size under Heterogeneity

Some Recent Results on Zelterman Estimation

Some simulation results
Main Conclusions

- Estimators of Chao and Zelterman closely related
- however: Zelterman Estimation offers more flexibility because ...
  - increasing its efficiency via truncated likelihood
  - incorporation of case data
  - incorporation of prior information by means of covariates
Generalizing the idea of Zelterman: truncation or censoring?

▶ disadvantage of conventional Zelterman: uses only $f_1$ and $f_2$
▶ idea of truncation: ignore all counts different from 1 and 2
▶ idea of censoring: use marginal likelihood for all counts of 2 and larger

\[ p_1 = P(Y = 1) = \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \lambda \]

\[ p_{2^+} = P(Y > 1) = \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} [\lambda^2 / 2! + \lambda^3 / 3! + \ldots] \]

\[ = 1 - \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \lambda \]
Generalizing the idea of Zelterman: truncation or censoring?

leads to the binomial likelihood

\[ f_1 \log(p_1) + f_{2+} \log(p_{2+}) \]

since

\[ \hat{p}_1 = \frac{f_1}{n} \]

we have

\[ \frac{f_1}{n} = \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \frac{\lambda}{\exp(\lambda) - 1} \]

\[ \approx \frac{1}{\lambda + \lambda^2/2} \]

leads to

\[ \hat{\lambda}_C = \frac{2(n - f_1)}{f_1} \]
frequently: evidence for a 2-component mixture model

Table: Amount of heterogeneity occurring in the data sets

<table>
<thead>
<tr>
<th>Example</th>
<th>Non-parametric mixture model</th>
</tr>
</thead>
<tbody>
<tr>
<td>McKendrick</td>
<td>homogeneity</td>
</tr>
<tr>
<td>Matthews</td>
<td>2-component</td>
</tr>
<tr>
<td>Scrapie</td>
<td>2-component</td>
</tr>
<tr>
<td>Drug Use L.A.</td>
<td>3-component</td>
</tr>
<tr>
<td>terrorist activity</td>
<td>6-component</td>
</tr>
</tbody>
</table>